A note on Talagrand's concentration inequality for empirical processes.

Dmitriy Panchenko
Department of Mathematics and Statistics
The University of New Mexico
panchenk@math.unm.edu
http://www.math.unm.edu/~panchenk/

January 2001

Abstract

In this paper we revisit Talagrand's proof of concentration inequality for empirical processes. We give a different shorter proof of the main technical lemma that garantees the existence of a certain kernel. Our proof provides the almost optimal value of the constant involved in the statement of this lemma.

1 Introduction and the proof of main lemma.

This paper was motivated by the Section 4 of the "New concentration inequalities in product spaces" by Michel Talagrand. For the most part we will keep the same notations with possible minor changes. We slightly weaken the definition of the distance $m(A,x)$ below compared to [10], but, essentially, this is what is used in the proof of the concentration inequality for empirical processes. The Theorem 1 below is to Theorem 4.2 in [10] and we assume that the reader is familiar with the proof. The main technical step, Proposition 4.2 in [10], is proved differently and constitutes the statement of Lemma 1 below.

Let Ω^n be a measurable product space with a product measure μ^n. Consider a probability measure ν on Ω^n and $x \in \Omega^n$. If $\mathcal{C}_i = \{y \in \Omega^n : y_i \neq x_i\}$, we consider the image of the restriction of μ to $\mathcal{C}_i$ by the map $y \to y_i$, and it's Radon-Nikodym derivative d_i with respect to μ. As in [10] we assume that Ω is finite and each point is measurable with a positive measure. Let m be a number of atoms in Ω and $p_1, \ldots, p_m$ be their probabilitites. By the definition of d_i we have

$$\int_{\mathcal{C}_i} g(y_i) d\mu(y) = \int_{\Omega} g(y_i) d_i(y_i) d\nu(y_i).$$

Consider the function

$$\psi(x) = \begin{cases} x^2/4, & \text{when } x \le 2, \\ x-1, & \text{when } x \ge 2. \end{cases}$$

We set

$$m(\nu, x) = \sum_{i \le n} \int \psi(d_i) d\mu \quad \text{and} \quad m(A, x) = \inf\{m(\nu, x) : \nu(A) = 1\}.$$

Theorem 1 *Let $L \ge 1.12$. Then for any n an $A \subseteq \Omega^n$ we have*

$$\int \exp \frac{1}{L} m(A, x) dP(x) \le \frac{1}{P(A)}. \tag{1.1}$$

As we mentioned above the proof is identical to [10] where Proposition 4.2 is substituted by the following lemma.

Lemma 1 *Let $g_1 \ge g_2 \ge \ldots \ge g_m > 0$. For $L \ge 1.12$ there exist $\{k_j^i : 1 \le j < i \le m\}$ such that*

$$k_j^i \ge 0, \quad \sum_{j<i} k_j^i p_j \le 1 \tag{1.2}$$

and

$$\sum_{i \le m} \frac{p_i}{g_i} \exp\Big\{\sum_{j<i}\Big(\log\frac{g_i}{g_j} k_j^i + \frac{1}{L}\psi(k_j^i)\Big)p_j\Big\} \le \frac{1}{p_1 g_1 + \ldots + p_m g_m}. \tag{1.3}$$

Remark: This lemma does not hold for $L \le 1.07$ (it is easy to construct the counterexample for $m = 2$), which means that $L = 1.12$ is close to the optimal.

Proof: The proof is by induction on the number of atoms m. The statement of lemma is trivial for $m = 1$. Note that in order to show the existence of $\{k_j^i\}$ in the statement of lemma one should try to minimize the left side of (1.3) with respect to $\{k_j^i\}$ under the constraints (1.2). Note also that each term on the left side of (1.3) has its own set of k_j^i, $j < i$ and, therefore, minimization can be performed for each term separately. We assume that k_j^i are chosen in an optimal way minimizing the left side of (1.3) and it will be convenient to take among all such optimal choices the one maximizing $\sum_{j<i} k_j^i p_j$ for all $i \le m$. To make the induction step we will start by proving the following statement, where we assume that k_j^i correspond to the specific optimal choice indicated above.

Statement. *For all $i \le m$, we have*

$$\sum_{j<i} k_j^i p_j < 1 \Longleftrightarrow \log\frac{g_1}{g_i} < \frac{1}{L} \text{ and } \sum_{j<i} 2L \log\frac{g_j}{g_i} p_j < 1. \tag{1.4}$$

In this case $k_j^i = 2L\log\frac{g_j}{g_i}$.

Proof: Let us fix i throughout the proof of the statement. We first assume that the left side of (1.4) holds. Suppose that $\log\frac{g_1}{g_i} \ge \frac{1}{L}$. In this case, since $\sup\{\psi'(x) : x \in \mathcal{R}\} \le 1$, one would decrease the left side of (1.3) by increasing k_1^i until $\sum_{j<i} k_j^i p_j = 1$ which contradicts the

choice of k_j^i. On the other hand, $\log \frac{g_1}{g_i} < \frac{1}{L}$ implies that $k_j^i \le 2$, since for $k \ge 2$, $\psi(k) = k-1$ and the choice of k_j^i would only increase the left side of (1.3). For $k \le 2$, $\psi(k) = k^2/4$ and

$$\operatorname{argmin}\Big(k \log \frac{g_i}{g_j} + \frac{k^2}{4L}\Big) = 2L \log \frac{g_j}{g_i}.$$

Hence, if $\sum_{j<i} 2L \log \frac{g_j}{g_i} p_j \ge 1$ then since $\sum_{j<i} k_j^i p_j < 1$ the set

$$\mathcal{J} := \{j : k_j^i \le 2L \log \frac{g_j}{g_i}\} \ne \emptyset$$

is not empty. But again this would imply that $\sum_{j<i} k_j^i p_j = 1$; otherwise, increasing k_j^i for $j \in \mathcal{J}$ would decrease the left side of (1.3). This completes the prove of the statement.

$\square$

(1.4) implies that if $\sum_{j<i} k_j^i p_j < 1$ then $\sum_{j<l} k_j^l p_j < 1$, for $l \le i$. Therefore, the equality $\sum_{j<m-1} k_j^{m-1} p_j = 1$ would imply $\sum_{j<m} k_j^m p_j = 1$. Let us first consider the case when $\sum_{j<m-1} k_j^{m-1} p_j = 1$. This step is meaningless for $m = 2$ and should simply be skipped. We will now show that $k_j^m = k_j^{m-1}$, $j < m-1$ and $k_{m-1}^m = 0$. Indeed,

$$\begin{aligned} &\inf_{\sum_{j<m} k_j p_j = 1} \sum_{j<m} \Big(\log \frac{g_m}{g_j} k_j + \frac{1}{L}\psi(k_j)\Big) p_j = \\ &= \log \frac{g_m}{g_{m-1}} + \inf_{\sum_{j<m} k_j p_j = 1} \Big(\sum_{j<m-1} \Big(\log \frac{g_{m-1}}{g_j} k_j + \frac{1}{L}\psi(k_j)\Big) p_j + \frac{1}{L}\psi(k_{m-1}) p_{m-1}\Big). \quad (1.5) \end{aligned}$$

Since $\sum_{j<m-1} k_j^{m-1} p_j = 1$ it is advantageous to set $k_{m-1}^m = 0$ and $k_j^m = k_j^{m-1}$, $j < m-1$. In this case

$$\frac{p_m}{g_m} \exp\Big\{\sum_{j<m} \Big(\log \frac{g_m}{g_j} k_j^m + \frac{1}{L}\psi(k_j^m)\Big) p_j\Big\} = \frac{p_m}{g_{m-1}} \exp\Big\{\sum_{j<m-1} \Big(\log \frac{g_{m-1}}{g_j} k_j^{m-1} + \frac{1}{L}\psi(k_j^{m-1})\Big) p_j\Big\}.$$

By induction assumption (1.3) holds for the sets $(g_1, \ldots, g_{m-1})$ and $(p_1, \ldots, p_{m-1} + p_m)$. Since $p_{m-1} g_{m-1} + p_m g_m \le (p_{m-1} + p_m) g_{m-1}$, it is clear that it holds for $(g_1, \ldots, g_m)$ and $(p_1, \ldots, p_m)$.

Now we will assume that $\sum_{j<m-1} k_j^{m-1} p_j < 1$ or, equivalently, $\log \frac{g_1}{g_{m-1}} < \frac{1}{L}$ and $\sum_{j<m-1} 2L \log \frac{g_j}{g_{m-1}} p_j < 1$. It is obvious that in this case there exist $g_0 < g_{m-1}$ such that for $g_m \in (g_0, g_{m-1}]$ both $\log \frac{g_1}{g_m} < \frac{1}{L}$ and $\sum_{j<m} 2L \log \frac{g_j}{g_m} p_j < 1$ hold and, therefore, $\sum_{j<m} k_j^m p_j < 1$. We assume that g_0 is the smallest number with such properties. Let us show that for a fixed $g_1, \ldots, g_{m-1}$ the case of $g_m < g_0$ can be converted to $g_m = g_0$. Indeed, take $g_m < g_0$. Clearly, $\sum_{j<m} k_j^m p_j = 1$. In this case (1.5) still holds and implies that k_j^m do not depend on g_m for $g_m < g_0$. We have

$$\frac{p_m}{g_m} \exp\Big\{\sum_{j<m} \Big(\log \frac{g_m}{g_j} k_j^m + \frac{1}{L}\psi(k_j^m)\Big) p_j\Big\} = \frac{p_m}{g_{m-1}} \exp\Big\{\sum_{j<m} \Big(\log \frac{g_{m-1}}{g_j} k_j^m + \frac{1}{L}\psi(k_j^m)\Big) p_j\Big\},$$

which means that for $g_m < g_0$ the left side of the inequality (1.3) does not depend on g_m. Since $(p_1 g_1 + \ldots + p_m g_m)^{-1}$ decreases in g_m it's enough to prove the inequality for $g_m = g_0$.

Hence, we can finally assume that $\log \frac{g_1}{g_m} \le \frac{1}{L}$, $\sum_{j<m} 2L \log \frac{g_j}{g_m} p_j \le 1$ and $k_j^i = 2L \log \frac{g_j}{g_i}$. (1.3) can be rewritten as

$$\sum_{i\le m} \frac{p_i}{g_i} \exp\Big\{-L \sum_{j<i} \Big(\log \frac{g_j}{g_i}\Big)^2 p_j\Big\} \le \frac{1}{p_1 g_1 + \ldots + p_m g_m}. \tag{1.6}$$

It is easy to see that by induction hypothesis (1.6) holds for $g_m = g_{m-1}$. To prove it for $g_m < g_{m-1}$ we will compare the derivatives of both sides of (1.6) with respect to g_m. It is enough to have

$$\frac{p_m}{g_m} \exp\Big\{-L \sum_{j<m} \Big(\log \frac{g_m}{g_j}\Big)^2 p_j\Big\} \Big(-\frac{1}{g_m} - 2L \sum_{j<m} \log \frac{g_m}{g_j} p_j \frac{1}{g_m}\Big) \ge -\frac{p_m}{(p_1 g_1 + \ldots + p_m g_m)^2}$$

or, equivalently,

$$\exp\Big\{-L \sum_{j<m} \Big(\log \frac{g_m}{g_j}\Big)^2 p_j\Big\} \Big(1 - 2L \sum_{j<m} \log \frac{g_j}{g_m} p_j\Big) \le \Big(\frac{g_m}{p_1 g_1 + \ldots + p_m g_m}\Big)^2.$$

Since $1 - x \le e^{-x}$ for $x \ge 0$ it's enough to show

$$\exp\Big\{-L \sum_{j<m} p_j \Big(\Big(\log \frac{g_j}{g_m}\Big)^2 + 2 \log \frac{g_j}{g_m}\Big)\Big\} \le \Big(\frac{g_m}{p_1 g_1 + \ldots + p_m g_m}\Big)^2.$$

One can check that $(\log x)^2 + 2 \log x$ is concave for $x \ge 1$. If we express $g_j = \lambda_j g_1 + (1 - \lambda_j) g_m$, $j = 1, \ldots, m-1$, then

$$\sum_{j<m} p_j \Big(\Big(\log \frac{g_j}{g_m}\Big)^2 + 2 \log \frac{g_j}{g_m}\Big) \ge (\sum_{j<m} p_j \lambda_j) \Big(\Big(\log \frac{g_1}{g_m}\Big)^2 + 2 \log \frac{g_1}{g_m}\Big)$$

$$p_1 g_1 + \ldots + p_m g_m = (\sum_{j<m} p_j \lambda_j) g_1 + (p_m + \sum_{j<m} (1 - \lambda_j) p_j) g_m.$$

If we denote $p = \sum_{j<m} p_j \lambda_j$ and $t = \log \frac{g_1}{g_m}$ we have to prove

$$\exp\Big\{-Lp(t^2 + 2t)\Big\} \le \Big(\frac{1}{pe^t + 1 - p}\Big)^2, \;\; 0 \le p \le 1, \, 0 \le t \le \frac{1}{L}. \tag{1.7}$$

Equivalently,

$$\varphi(p,t) = (pe^t + 1 - p) \exp\Big\{-\frac{L}{2} p(t^2 + 2t)\Big\} \le 1, \;\; 0 \le p \le 1, \, 0 \le t \le \frac{1}{L}.$$

We have

$$\varphi_t'(p,t) = \varphi(p,t) \Big(\frac{pe^t}{pe^t + 1 - p} - Lp(t+1)\Big).$$

Since for all $p > 0$ $\varphi(p,0) = 1$ we need $\varphi_t'(p,0) = p(1 - L) \le 0$, or $L \ge 1$, which holds if $L \ge 1.12$. It is easy to see that $\varphi_t'(p,t) = 0$ in at most one point t. In combination with

$\varphi_t'(p,0) \le 0$ it implies that for a fixed p maximum of $\varphi(p,t)$ is attained at $t=0$ or $t=1/L$. Therefore, we have to show $\varphi(p,1/L) \le 1$, $0 \le p \le 1$. We have,

$$\varphi_p'(p,\frac{1}{L}) = \varphi(p,\frac{1}{L})\Big(\frac{e^{\frac{1}{L}}-1}{pe^{\frac{1}{L}}+1-p} - \frac{L}{2}\Big(\frac{1}{L^2}+2\frac{1}{L}\Big)\Big).$$

Since $\varphi(0,\frac{1}{L}) = 1$ we should have $\varphi_p'(0,\frac{1}{L}) \le 0$ which would also imply $\varphi_p'(p,\frac{1}{L}) \le 0$, $p > 0$. One can check that

$$\varphi_p'(0,\frac{1}{L}) = e^{\frac{1}{L}} - 1 - \frac{1}{2}\Big(\frac{1}{L}+2\Big) < 0$$

for $L \ge 1.12$. This finishes the proof of Lemma.

2 One concentration inequality for empirical processes.

Given Theorem 1 one can proceed as in [10] to obtain the classical form of concentration inequality for the empirical process around its mean.

We will now show that in one special case which allows certain simplifications the technique of Talagrand allows to obtain rather sharp concentration result with explicit constants. Consider the countable class of measurable functions $\mathcal{F} = \{f : \Omega \to [0,1]\}$. Consider the following function on Ω^n

$$Z(x) = \sup_{f\in\mathcal{F}} \sum_{i\le n}(\mu f - f(x_i))$$

where $\mu f := \int f d\mu$. It often happens in applications (see [3], [4], [5]), especially in the case when the empirical process is defined over the family of sets, that the uniform variance $n\sup_{f\in\mathcal{F}} \mathrm{Var} f$ is simply bounded by uniform second moment

$$\sigma^2 = n \sup_{f\in\mathcal{F}} \mu f^2 \tag{2.1}$$

for which one has an apriori bound. Talagrand's technique gives in this case a proof of the following concentration inequalities.

Theorem 2 *Let $L = 1.12$ and M be a median of Z. Then*

$$\mathbb{P}(Z \ge M + 2\max(Lu, \sigma\sqrt{Lu})) \le 2e^{-u},$$

$$\mathbb{P}(Z \le M - 2\max(Lu, \sigma\sqrt{Lu})) \le 2e^{-u}.$$

Proof. Without loss of generality we assume that $\mathcal{F}$ is finite. Given a let us consider the set $A = \{Z(x) \le a\}$. For a fixed x let $f \in \mathcal{F}$ be such that

$$Z(x) = \sum_{i\le n}(\mu f - f(x_i)). \tag{2.2}$$

Then for the probability measure ν such that $\nu(A) = 1$ we have

$$Z(x) - a \leq \int \Big(\sum_{i \leq n} (\mu f - f(x_i)) - \sum_{i \leq n} (\mu f - f(y_i)) \Big) d\nu(y)$$
$$= \sum_{i \leq n} \int (f(y_i) - f(x_i)) d_i(y_i) d\mu(y_i) \leq \sum_{i \leq n} \int f(y_i) d_i(y_i) d\mu(y_i).$$

As is easily checked for $v \geq 0$, and $0 \leq u \leq 1$,

$$uv \leq u^2 + \psi(v).$$

Therefore, for any $\delta \geq 1$

$$Z(x) - a \leq \delta \sum_{i \leq n} \int \frac{f(y_i)}{\delta} d_i(y_i) d\mu(y_i) \leq \frac{1}{\delta}\sigma^2 + \delta \sum_{i \leq n} \int \psi(d_i) d\mu$$

Taking the infimum over ν we get

$$Z(x) \leq a + \frac{1}{\delta}\sigma^2 + \delta m(A, x).$$

Theorem 1 then implies that for $L = 1.12$ with probability at least $1 - \frac{1}{P(Z \leq a)} e^{-u}$

$$Z(x) \leq a + 2\max(Lu, \sigma\sqrt{Lu}).$$

Applied to $a = M$ - median of Z, and to $a = M - 2\max(Lu, \sigma\sqrt{Lu})$ gives the result.

Remark. It is interesting to notice that the bounds of Theorem 2 seem to avoid the "singular" behaviour of the general bounds expressed in terms of the weak variance $n \sup_{f \in \mathcal{F}} \mathrm{Var} f$ (see [6]), when the linear dependance of the term $(1+\varepsilon)M$ on ε reguires the factor of the order ε^{-1} in the last term of the bound $\varepsilon^{-1}u$. Under the assumptions of Theorem 2, one can also avoid this "singularity" using the recent result of Emmanuel Rio [8], that provides rather sharp constants too, and the concentration is around mean instead of median.

Acknowledgments. We want to thank Michel Talagrand for pointing out the recent results of Emmanuel Rio.

References

[1] Dembo, A., Information inequalities and concentration of measure, *Ann. Probab.,* **25** (1997), 527-539.

[2] Ledoux, M., Talagrand deviation inequalities for product measures, *ESAIM: Probab. Statist.*, **1** (1996), 63 - 87.

[3] Koltchinskii, V., Panchenko, D. Rademacher processes and bounding the risk of function learning, *High Dimensional Probability II,* Eds. E.Giné, D.Mason and J.Wellner, (2000), pp. 443 - 459.

[4] Koltchinskii, V., Panchenko, D., Empirical margin distributions and bounding the generalization error of combined classifiers, submitted, (2000)

[5] Koltchinskii, V., Panchenko, D., Lozano, F., Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins, submitted, (2000).

[6] Massart, P., About the constants in Talagrand's concentration inequalities for empirical processes, *Ann. Probab.,* **28** (2000), 863 - 885.

[7] Boucheron, S., Lugosi, G., Massart, P., A sharp concentration inequality with applications, Prépublications 99-25 du laboratoire de mathématiques, Université Paris-Sud, (1999)

[8] Rio E., Inégalités exponentielles pour les processus empiriques, *C.R. Acad. Sci. Paris,* t.330, Série I (2000), 597-600.

[9] Talagrand, M., Concentration of measure and isoperimetric inequalities in product spaces, *Publications Mathématiques de l'I.H.E.S.* **81** (1995), 73-205.

[10] Talagrand, M., New concentration inequalities in product spaces. *Invent. Math.,* **126** (1996) , 505-563.